Machine Learning

Homework 2

Jiale Shi

February 16, 2019

1 Linear Regression

Consider a data set in which each data point y_n is associated with a weighting factor $r_n > 0$. Therefore, the sum of square error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{y_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 \tag{1}$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (a) data dependent noise variance (b) replicated data points.
Solution:

$$\begin{aligned} &\mathbf{w}^* = \text{argmin}_{\mathbf{w}} E_D(\mathbf{w}) \\ &0 = \frac{dE_D(\mathbf{w}^*)}{d\mathbf{w}^*} = -\sum_{n=1}^{N} r_n\phi(\mathbf{x}_n)\{y_n - \mathbf{w}^{*T}\phi(\mathbf{x}_n)\} \\ &\sum_{n=1}^{N} r_n\phi(\mathbf{x}_n)\{y_n - \mathbf{w}^{*T}\phi(\mathbf{x}_n)\} = 0 \end{aligned} \tag{2}$$

In terms of (a) data dependent noise variance: $E_D(\mathbf{w})$ scales the noise of data (β).
In terms of (b) replicated data points: It shows how many times does each data point is observed.

2 Evidence

A. In linear regression, the marginal likelihood function is given as

$$p(t|\alpha,\beta) = \int p(t|\omega,\beta)p(\omega|\alpha)d\omega \tag{3}$$

where

$$p(\omega|\alpha) = \mathcal{N}(\omega|0,\alpha^{-1}I) \tag{4}$$

$$\log p(t|\omega,\beta) = \sum_{n=1}^{N} \log \mathcal{N}(t_n|\omega^T\phi(\mathbf{x}_n),\beta^{-1}) \tag{5}$$

$$E_D(\omega) = \frac{1}{2}\sum_{n=1}^{N}\{t-\omega^T\phi(\mathbf{x}_n)\}^2 \tag{6}$$

Using the above equations, derive an expression for the marginal likelihood.
Solution:

$$\begin{aligned}
&\log p(t|\omega,\beta) = \sum_{n=1}^{N} \log \mathcal{N}(t_n|\omega^T\phi(\mathbf{x}_n),\beta^{-1}) \\
&p(t|\omega,\beta) = \frac{1}{(2\pi)^{N/2}}\beta^{N/2}e^{-\beta E_D} \\
&E_D(\omega) = \frac{1}{2}\sum_{n=1}^{N}\{t-\omega^T\phi(\mathbf{x}_n)\}^2
\end{aligned} \tag{7}$$

$$\begin{aligned}
p(t|\alpha,\beta) &= \int p(t|\omega,\beta)p(\omega|\alpha)d\omega \\
&= \int \frac{1}{(2\pi)^{N/2}}\beta^{N/2}e^{-\beta E_D}\mathcal{N}(\omega|0,\alpha^{-1}I)d\omega \\
&= (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\int \exp\{-E(\omega)\}d\omega \\
&= (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}e^{-E(\mathbf{m}_N)}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\mathbf{A} = \alpha\mathbf{I} + \beta\phi^T\phi \\
&E(\mathbf{m}_N) = \frac{\beta}{2}||\mathbf{t}-\phi\mathbf{m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \\
&\mathbf{m}_N = \beta\mathbf{A}^{-1}\phi^T\mathbf{t}
\end{aligned} \tag{9}$$

B. The conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distribution. This property also holds for the case of the conditional Gaussian distribution $p(t|x,\omega,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|\omega^T\phi(x_n),\beta^{-1})$ of the linear regression model. If we consider the likelihood function

$$p(t|x,\omega,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|\omega^T\phi(x_n),\beta^{-1}) \tag{10}$$

then the conjugate prior for ω and β is given by

$$p(\omega, \beta) = \mathcal{N}(\omega|m_0, \beta^{-1}S_0)\text{Gam}(\beta|a_0, b_0) \tag{11}$$

(a) Show that the corresponding posterior distribution takes the same functional form, so that

$$p(\omega, \beta|t) = \mathcal{N}(\omega|m_N, \beta^{-1}S_N)\text{Gam}(\beta|a_N, b_N) \tag{12}$$

and find expressions for the posterior parameters $\Updownarrow_N$, S_N, a_N and b_N.
Solution:

$$\begin{aligned}
p(\omega, \beta|t) &= p(t|x, \omega, \beta)p(\omega, \beta) \\
&= \left\{\prod_{n=1}^{N} \mathcal{N}(t_n|\omega^T\phi(x_n), \beta^{-1})\right\} \mathcal{N}(\omega|m_0, \beta^{-1}S_0)\text{Gam}(\beta|a_0, b_0) \\
&= \frac{1}{(2\pi)^{N/2}}(\frac{1}{\beta})^{-N/2} \exp\left(-\beta\frac{(\mathbf{y}-\mathbf{\Phi w})^T(\mathbf{y}-\mathbf{\Phi w})}{2}\right) \\
&\cdot \frac{b_0^{a_0}}{(2\pi)^{D/2}|\mathbf{S}_0|^{1/2}\Gamma(a_0)}(\frac{1}{\beta})^{-(a+D/2+1)} \exp\left(-\frac{(\mathbf{w}-m_0)^T(\mathbf{S}_0)^{-1}(\mathbf{w}-m_0)+2b}{2\sigma^2}\right) \\
&= \frac{b_0^{a_0}}{(2\pi)^{(N+D)/2}|\mathbf{S}_0|^{1/2}\Gamma(a_0)}(\frac{1}{\beta})^{-(a+(D+N)/2+1)} \\
&\cdot \exp\left(-\beta\frac{(\mathbf{w}-m_0)^T(\mathbf{S}_0)^{-1}(\mathbf{w}-m_0)+(\mathbf{y}-\mathbf{\Phi w})^T(\mathbf{y}-\mathbf{\Phi w})+2b_0}{2}\right)
\end{aligned} \tag{13}$$

Let us define the following:

$$\begin{aligned}
\mathbf{S}_N &= (\mathbf{S}_0^{-1} + \mathbf{\Phi}^\mathbf{T}\mathbf{\Phi})^{-1} \\
\mathbf{m}_N &= \mathbf{V}_N(\mathbf{S}_0^{-1}m_0 + \mathbf{\Phi}^\mathbf{T}\mathbf{y}) \\
a_N &= a_0 + N/2 \\
b_N &= b_0 + \frac{1}{2}\left(m_0\mathbf{S}_0^{-1}m_0 + \mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{m}_N\mathbf{S}_N^{-1}\mathbf{m}_N\right)
\end{aligned} \tag{14}$$

with these definitions,

$$\begin{aligned}
p(\mathbf{w}, \sigma^2|D) &\propto (\frac{1}{\beta})^{-(a_N+D/2+1)} \exp\left(-\beta\frac{(\mathbf{w}-\mathbf{m}_N)^T(S_N)^{-1}(\mathbf{w}-\mathbf{m}_N)^T+2b_N}{2}\right) \\
&= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\text{Gam}(\beta|a_N, b_N)
\end{aligned} \tag{15}$$

(b) Show that the marginal probability of the data (model evidence) is given by

$$p(t) = \frac{1}{2\pi^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|S_N|^{1/2}}{|S_0|^{1/2}}\frac{\Gamma(a_N)}{\Gamma(a_0)} \tag{16}$$

Solution:

$$p(t) = \int \int p(t|x, \omega, \beta) p(\omega, \beta) dw d\beta \tag{17}$$

3 Mixture of conjugate priors

A. Show that a mixture of conjugate priors is indeed a conjugate prior.
Solution: From Murphy's book, 5.4.4 mixture of conjugate priors. We can represent a mixture of conjugate priors by introducing a latent indicator variable z, where $z = k$ means that θ comes from mixture component k. The prior has the form

$$p(\theta) = \sum_{k=1}^{m} p(z = k)p(\theta|z = k) \tag{18}$$

Then the posterior can be written as

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{19}$$

Combine the above two equations

$$p(\theta|D) = \sum_{k=1}^{m} \frac{1}{p(D)} p(z = k)p(\theta|z = k)p(D|\theta) \tag{20}$$

And, if we know θ, no matter what the z is, $p(D|\theta)$ won't change. Therefore, $p(D|\theta) = p(D|\theta, z = k)$.

$$\begin{aligned} p(\theta|D) &= \sum_{k=1}^{m} \frac{p(z = k)}{p(D)} p(\theta|z = k)p(D|\theta, z = k) \\ &= \sum_{k=1}^{m} \frac{p(D|z = k)p(z = k)}{p(D)} \frac{p(\theta|z = k)p(D|\theta, z = k)}{p(D|z = k)} \\ &= \sum_{k=1}^{m} \frac{p(D|z = k)p(z = k)}{p(D)} p(\theta|D, z = k) \\ &= \sum_{k=1}^{m} p(z = k|D)p(\theta|D, z = k) \end{aligned} \tag{21}$$

$p(z = k|D)$ is the posterior weight. $\sum_{k=1}^{m} p(z = k|D) = 1$
B. Suppose we use the mixture prior $p(\theta) = 0.5\text{Beta}(\theta|a_1, b_1)+0.5\text{Beta}(\theta|a_2, b_2)$, where $a_1 = b_1 = 20$, $a_2 = b_2 = 10$ and we observe N_1 head and N_0 tails. Derive an expression and write a computer code for evaluating the posterior. Consider $N_1 = 20$ heads and $N_0 = 10$ tails. Show with a plot a comparison between the prior and the posterior.

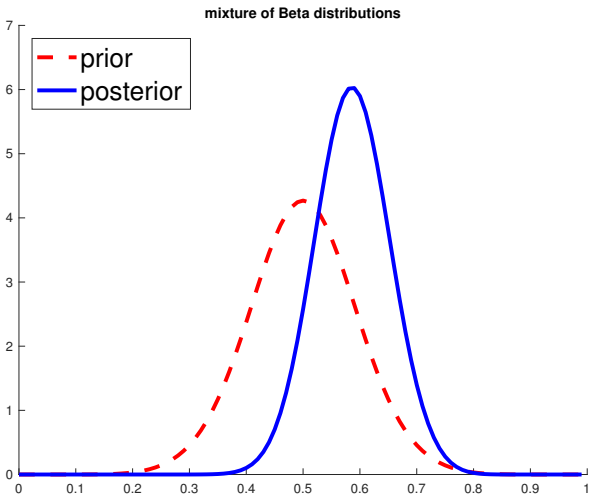

Figure 1: Comparison between the prior and the posterior

4 Optimal threshold on classification probability

Consider a case where we have learned a conditional probability distribution $p(y|x)$. Suppose there are only two classes, and let $p_0 = p(y = 0|x)$ and $p_1 = p(y = 1|x)$. Consider the loss matrix shown in Table 1.

A. Show that the decision $\hat{y}$ that minimizes the expected loss is equivalent to setting a probability threshold θ and predicting $\hat{y} = 0$ if $p_1 < \theta$ and $\hat{y} = 1$ if $p_1 \geq \theta$. Derive θ as a function of λ_{01} and λ_{10}.

Solution:

$$\begin{aligned} p_0 &= p(y = 0|\mathbf{x}) \\ p_1 &= p(y = 1|\mathbf{x}) \\ p_0 + p_1 &= 1 \\ p_0 &= 1 - p_1 \end{aligned} \tag{22}$$

predicting $\hat{y} = 0$ if $p_1 < \theta$ and $\hat{y} = 1$ if $p_1 > \theta$
$L_{FN} = \lambda_{01}$ and $L_{FP} = \lambda_{10}$

One should pick the class $\hat{y} = 1$ if

$$\begin{aligned} \frac{p_1}{p_0} &> \frac{\lambda_{10}}{\lambda_{01}} \\ \frac{p_1}{1-p_1} &> \frac{\lambda_{10}}{\lambda_{01}} \\ p_1 &> \frac{\frac{\lambda_{10}}{\lambda_{01}}}{1+\frac{\lambda_{10}}{\lambda_{01}}} \end{aligned} \tag{23}$$

Therefore,

$$\theta = \frac{\frac{\lambda_{10}}{\lambda_{01}}}{1+\frac{\lambda_{10}}{\lambda_{01}}} \tag{24}$$

B. Derive the loss function where the threshold is 0.1.
Solution:
From part **A**

$$0.1 = \theta = \frac{\frac{\lambda_{10}}{\lambda_{01}}}{1+\frac{\lambda_{10}}{\lambda_{01}}} \tag{25}$$

$$\frac{\lambda_{10}}{\lambda_{01}} = \frac{1}{9} \tag{26}$$

if $p_1 < 0.1$, $\hat{y} = 0$

$$L = p_1 \lambda_{01} \tag{27}$$

and if $p_1 > 0.1$, $\hat{y} = 1$

$$L = p_0 \lambda_{10} \tag{28}$$

5 Bayes Factor

A. Suppose we toss a coin N times and observe N_1 heads. Let $N_1 \sim \text{Bin}(N, \theta)$ and $\theta \sim \text{Beta}(1, 1)$. Show that the marginal likelihood is $p(N_1|N) = \frac{1}{N+1}$.
Solution:

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \\ &= \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right] \end{aligned} \tag{29}$$

$$\frac{1}{B(a+N_1, b+N_0)} = \frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}\binom{N}{N_1} \tag{30}$$

$$\begin{aligned} p(N_1|N) = p(\mathcal{D}) &= \binom{N}{N_1}\frac{B(a+N_1, b+N_0)}{B(a,b)} \\ &= \binom{N}{N_1}\frac{B(1+N_1, 1+N_0)}{B(1,1)} \quad = \frac{N!}{N_1!N_0!}\frac{\frac{N_1!N_0!}{(N+1)!}}{\frac{0!0!}{1!}} \\ &= \frac{1}{N+1} \end{aligned} \tag{31}$$

B. Suppose we toss a coin $N = 10$ times and observe $N_1 = 9$ heads. Let the null hypothesis be that the coin is fair, and the alternative be that the coin can have any bias, so $p(\theta) = \mathcal{U}(0, 1)$. Derive the Bayes factor $BF_{0,1}$ in favor of the biased coin hypothesis. What if $N = 100$ and $N_1 = 90$?
Solution:
If the coin is fair

$$P_0(N_1|N) = \binom{N}{N_1}(0.5)^{N_1}(0.5)^{N_0} = \binom{N}{N_1}(0.5)^N \tag{32}$$

If the coin can have any bias, from part A

$$p_1(N_1|N) = p(\mathcal{D}) = \frac{1}{N+1} \tag{33}$$

$$BF_{0,1} = \frac{P_0(N_1|N)}{\frac{1}{N+1}} = \binom{N}{N_1}(0.5)^N(N+1) \tag{34}$$

Therefore, if we toss a coin $N = 10$ times and observe $N_1 = 9$ heads,

$$BF_{0,1} = \binom{10}{9}(0.5)^{10}(10+1) = \frac{55}{512} = 0.1074 \tag{35}$$

Therefore, if we toss a coin $N = 100$ times and observe $N_1 = 90$ heads,

$$BF_{0,1} = \binom{100}{90}(0.5)^{100}(100+1) = 1.379 \times 10^{-15} \tag{36}$$

6 Behavior of training set error with increasing sample size, Multi-out regression and Ridge regression

A. The error on the test will always decrease as we get more training data, since the model will be better estimated. However, for sufficiently complex models, the error on the training set can increase as we get more training data, until we reach some plateau. Explain why.

Solution:

When the training data is not enough, these complex models are easy to be overfitting, which would leads to very low training error and very high test error.

For example, the true function is very complex. But when there are only two points, we get a linear model and the training error is 0 while the test error is very large. However, when we get more training data, the model begins to learn from the training data and becomes more and more close to the true function. As a result, the training error and test error would converge to one same plateau. Therefore, as we get more training data, the training error can increase and the test error would decrease.

B. When we have multiple independent outputs in linear regression, the model becomes

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}) = \prod_{j=1}^{M} \mathcal{N}(y_j|w_j^T \mathbf{x}_i, \sigma_j^2) \tag{37}$$

Since the likelihood factorizes across dimensions, so does the MLE. Thus,

$$\hat{W} = [\hat{w}_1, ..., \hat{w}_M] \tag{38}$$

where $\hat{w}_j = (X^T X)^{-1} X^T Y_{:,j}$.

In this exercise, we apply this result to a model with 2 dimensional response vector $y_i \in \mathbb{R}^2$. Suppose we have some binary input data $x_i \in \{0, 1\}$. The training data is as follows. Let us embed each x_i into 2d using the following basis function

$$\phi(0) = (1, 0)^T, \phi(1) = (0, 1)^T, \tag{39}$$

The model becomes

$$\hat{\mathbf{y}} = W^T \phi(x) \tag{40}$$

where W is a 2×2 matrix. Compute the MLE for W from the above data.

Solution:

$$\begin{aligned} &\hat{\mathbf{y}} = W^T \phi(x) \\ &y_j = \phi(x)^T \mathbf{w}_j \\ &y_1 = \phi(x)^T \mathbf{w}_1 \quad y_2 = \phi(x)^T \mathbf{w}_2 \end{aligned} \tag{41}$$

$$\mathbf{w}_1 = \begin{bmatrix} a \\ b \end{bmatrix} \quad \mathbf{w}_2 = \begin{bmatrix} c \\ d \end{bmatrix} \tag{42}$$

Using the least squares to minimize the sum squares error function L_s.

$$\begin{aligned} L_s =& (-1-a)^2 + (-1-c)^2 \\ &+ (-1-a)^2 + (-2-c)^2 \\ &+ (-2-a)^2 + (-1-c)^2 \\ &+ (1-b)^2 + (1-d)^2 \\ &+ (1-b)^2 + (2-d)^2 \\ &+ (2-b)^2 + (1-d)^2 \end{aligned} \tag{43}$$

when $a = -\frac{4}{3}$, $b = \frac{4}{3}$, $c = -\frac{4}{3}$, $d = \frac{4}{3}$, L_s gets the minimum value. There,

$$\begin{aligned} &\mathbf{w}_1 = \begin{bmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{bmatrix} \qquad\qquad \mathbf{w}_2 = \begin{bmatrix} -\frac{4}{3} \\ \frac{4}{3} \end{bmatrix} \\ &\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2] = \begin{bmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{bmatrix} \end{aligned} \tag{44}$$

C. Assume that $\hat{x} = 0$, so that the input data is centered. Show that the optimizer (in case of ridge regression) of

$$J(\mathbf{w}, w_0) = (\mathbf{y} - X\mathbf{w} - w_0\mathbf{1})^T(\mathbf{y} - X\mathbf{w} - w_0\mathbf{1}) + \lambda\mathbf{w}^T\mathbf{w} \tag{45}$$

is

$$w_0 = \bar{y} \tag{46}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{47}$$

Solution:

$$\begin{aligned} 0 = \frac{dJ}{dw_0} &= 2Nw_0 + 2(\mathbf{X}\mathbf{w} - \mathbf{y})^T\mathbf{1} \\ w_0 &= \bar{y} - \mathbf{W}\hat{x} \end{aligned} \tag{48}$$

Since the problem assume that $\hat{x} = 0$, therefore, $\mathbf{W}\hat{x} = 0$

$$w_0 = \bar{y} \tag{49}$$

$$0 = \frac{dJ}{d\mathbf{w}} = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\lambda\mathbf{w} - \mathbf{X}^T\mathbf{X} + 2w_0\mathbf{X}^T\mathbf{1} \tag{50}$$

Since the problem assume that $\hat{x} = 0$, therefore, $\mathbf{X}^T\mathbf{I} = 0$

$$\begin{aligned} 0 &= 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{I}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} \\ \mathbf{w} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{51}$$

D. For the data set provided, we want to fit a linear regression model with polynomial order M. In this regard, perform the following tasks:
(a) Compute the unknown coefficients based on MLE with $M = 2, 4, 10, 14$. Compute and plot the mean square error for the training and the test set corresponding to various polynomial orders.
Solution:

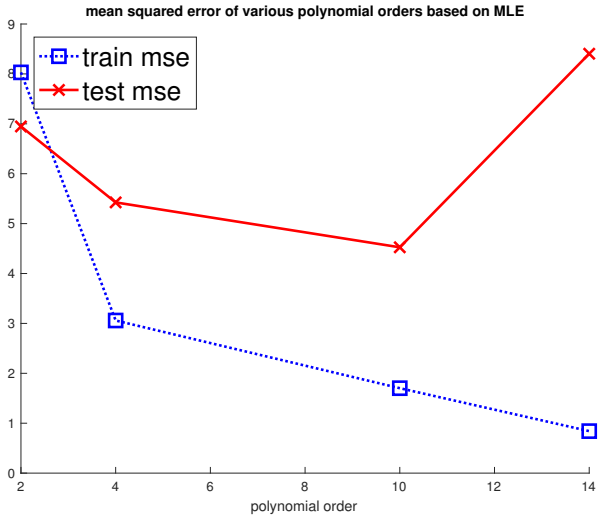

Figure 2: Mean Square Error for the training and test set corresponding to various polynomial orders based on MLE

From the results, we find that when polynomial order is 10, the mean square error for the test set is smallest.
(b) Compute the known coefficients based on ridge regression and plot the fitted function. Report the mean square error corresponding to the training and the test set.
Solution:
We set polynomial order is 10 and compute the unknown coefficients based on ridge regression. The fitted functions corresponding to different lambdas are Instead of reporting the mean square error, it is clearer to show the mean square error in the figure.
When $\log(\lambda) = -14.3528$, the test error is smallest.

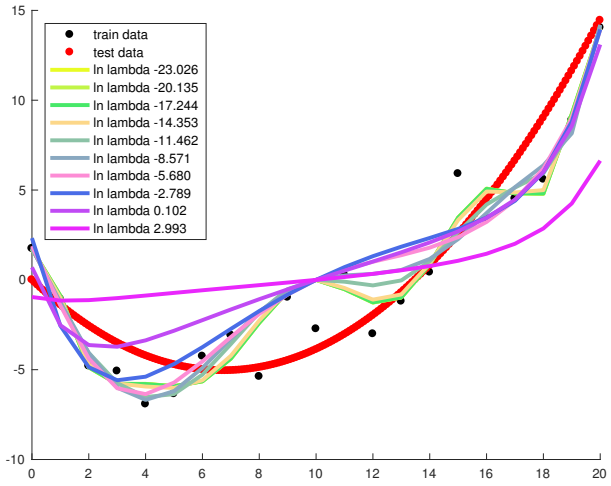

Figure 3: The fitted functions corresponding to different lambdas

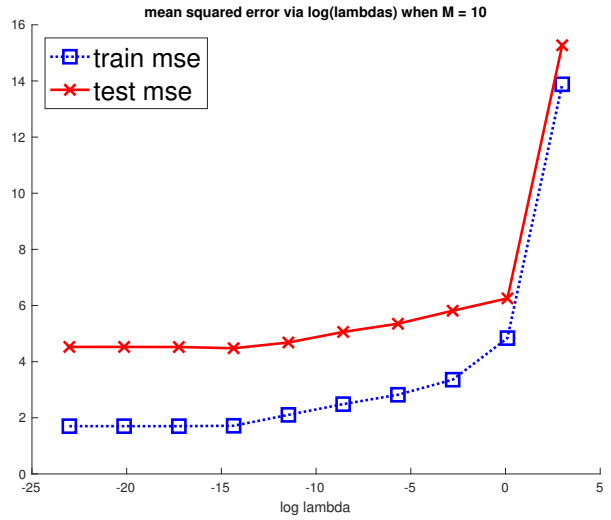

Figure 4: Mean Square Error for the training and test set corresponding to various lambdas with polynomial order is 10 based on redge regression

7 Bayesian linear regression

We consider a Bayesian linear regression model to fit a set of points x^i,y^i,$i = 1, ..., N$ data points with up to order $M = 5$ polynomials. For your implementa-

tion consider the data provided. The data is generated using the following data generation algorithm:
$x = 10\text{rand}(\text{N}, 1)$; generate input points
noise = 3randn(N,1); generate Gaussian noise
$y = (x-4)^2 + \text{noise}$; actual truth function
The particulars of the regression model are given below:

$$p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I}_N) \tag{52}$$

where $\mathbf{\Phi}$ is the design matrix. The prior to be considered is as follows:

$$p(\mathbf{w}|\sigma^2, \mathbf{\Phi}) = \mathcal{N}(0, \gamma\sigma^2\mathbf{I}) \tag{53}$$

where $p(\sigma^2) = \text{InvGamma}(a, b)$. The particular parameter that you should use in your implementation are $a = 0.1$, $b = 0.00001$ and $\gamma = 0.001/N$
A. Derive an expression for the posterior of $p(\mathbf{w}, \sigma^2|D)$, the marginal posterior $p(\mathbf{w}|D)$, predictive distribution $p(y|x, D)$ and the model evidence $p(D)$.
Solution:

$$\begin{aligned}
&p(\mathbf{w}, \sigma^2|D) = p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \sigma^2)(p(\mathbf{w}|\sigma^2, \mathbf{\Phi}) \cdot p(\sigma^2)) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I}_N)\mathcal{N}(0, \gamma\sigma^2\mathbf{I})\text{InvGamma}(a, b) \\
&= \frac{1}{(2\pi)^{N/2}}(\sigma^2)^{-N/2}\exp\left(-\frac{(\mathbf{y}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{y}-\mathbf{\Phi}\mathbf{w})}{2\sigma^2}\right) \\
&\cdot \frac{b^a}{(2\pi)^{D/2}|\gamma I|^{1/2}\Gamma(a)}(\sigma^2)^{-(a+D/2+1)}\exp\left(-\frac{\mathbf{w}^T(\gamma I)^{-1}\mathbf{w} + 2b}{2\sigma^2}\right) \\
&= \frac{b^a}{(2\pi)^{(N+D)/2}|\gamma I|^{1/2}\Gamma(a)}(\sigma^2)^{-(a+(D+N)/2+1)}\exp\left(-\frac{\mathbf{w}^T(\gamma I)^{-1}\mathbf{w} + (\mathbf{y}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{y}-\mathbf{\Phi}\mathbf{w}) + 2b}{2\sigma^2}\right)
\end{aligned} \tag{54}$$

Let us define the following:

$$\begin{aligned}
&\mathbf{V}_N = ((\gamma I)^{-1} + \mathbf{\Phi}^\mathbf{T}\mathbf{\Phi})^{-1} \\
&\mathbf{w}_N = \mathbf{V}_N((\gamma I)^{-1}w_0 + \mathbf{\Phi}^\mathbf{T}\mathbf{y}) = \mathbf{V}_N(\mathbf{\Phi}^\mathbf{T}\mathbf{y}) \\
&a_N = a + N/2 \\
&b_N = b + \frac{1}{2}(w_0(\gamma I)^{-1}w_0 + \mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{w}_N^T\mathbf{V}_N^{-1}\mathbf{w}_N) \\
&\quad = b + \frac{1}{2}(\mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{w}_N^T\mathbf{V}_N^{-1}\mathbf{w}_N)
\end{aligned} \tag{55}$$

with these definitions,

$$p(\mathbf{w}, \sigma^2|D) \propto (\sigma^2)^{-(a+(D+N)/2+1)}\exp\left(-\frac{(\mathbf{w}-\mathbf{w}_N)^T(V_N)^{-1}(\mathbf{w}-\mathbf{w}_N)^T + 2b}{2\sigma^2}\right) \tag{56}$$

$$
\begin{aligned}
p(\mathbf{w}|D) &\propto \int_0^\infty \propto (\sigma^2)^{-(a+(D+N)/2+1)} \exp\left(-\frac{(\mathbf{w}-\mathbf{w}_N)^T(V_N)^{-1}(\mathbf{w}-\mathbf{w}_N)^T + 2b}{2\sigma^2}\right) d(\sigma)^2 \\
&\propto \left[1 + \frac{(\mathbf{w}-\mathbf{w}_N)^T(V_N)^{-1}(\mathbf{w}-\mathbf{w}_N)}{2b_N}\right]^{-\frac{2a_N+D}{2}} \\
&= \mathcal{T}_D(\mathbf{w}, \frac{b_N}{a_N}\mathbf{V_N}, 2a_N)
\end{aligned} \tag{57}
$$

$$p(y|x, D) = \mathcal{T}_m(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \frac{b_N}{a_N}(\mathbf{I}_m + \mathbf{\Phi}\mathbf{V}_N\mathbf{\Phi}^T), 2a_N) \tag{58}$$

$$p(D) = \frac{1}{(2\pi)^N} \frac{|\mathbf{V}_N|^{1/2}}{|\gamma \mathbf{I}|^{1/2}} \frac{b^a}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a)} \tag{59}$$

B. For polynomial orders $M = 1, 2, 3, 4, 5$ plot the predictive mean and the predictive error bars. Your plots should also indicate the exact function as well as the training data points.

Solution:

$$\begin{aligned}
&\mathbf{V}_N = ((\gamma I)^{-1} + \mathbf{\Phi^T \Phi})^{-1} \\
&\mathbf{w}_N = \mathbf{V}_N((\gamma I)^{-1} w_0 + \mathbf{\Phi^T y}) = \mathbf{V}_N(\mathbf{\Phi^T y}) \\
&a_N = a + N/2 \\
&b_N = b + \frac{1}{2}(w_0(\gamma I)^{-1} w_0 + \mathbf{y^T y} - \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N) \\
&\quad = b + \frac{1}{2}(\mathbf{y^T y} - \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N)
\end{aligned} \tag{60}$$

the predictive mean is $\mathbf{\Phi w}_N$
the predictive variance matrix is $\frac{b_N}{a_N}(\mathbf{I}_m + \mathbf{X V}_N \mathbf{X}^T)$
the predictive error bar is the square root of diagonal value of the predictive variance matrix.

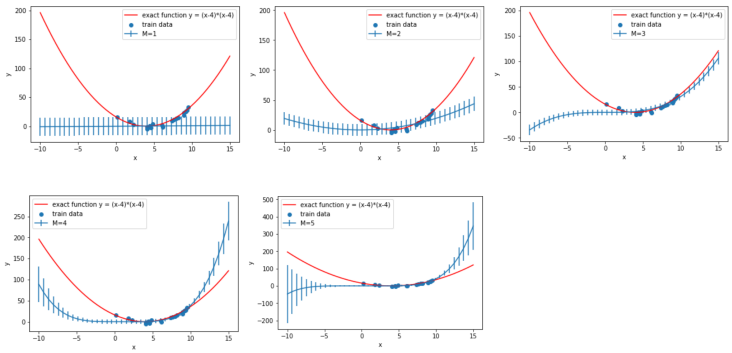

Figure 5: Predictive mean and the predictive error bars for polynomials $M = 1, 2, 3, 4, 5$

C. Draw samples of $\mathbf{w}$ and for each of them show the predictive mean. What additional information this graph provides that is not given in your plots in B above?

Solution:

$$
\begin{aligned}
p(\mathbf{w}|D) &= \mathcal{T}_D(\mathbf{w}, \frac{b_N}{a_N}\mathbf{V_N}, 2a_N) \\
&\propto \left[1 + \frac{(\mathbf{w} - \mathbf{w}_N)^T (V_N)^{-1} (\mathbf{w} - \mathbf{w}_N)}{2b_N}\right]^{-\frac{2a_N + D}{2}}
\end{aligned} \tag{61}
$$

Therefore, we need to draw samples from multi-variable Student's T distribution.

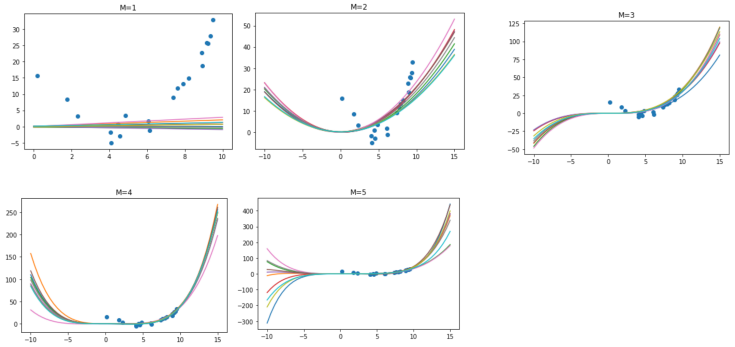

Figure 6: Draw samples of $\mathbf{w}$ from Student T distribution $M = 1, ..., 5$

D. Show a plot comparing the model evidence for polynomials $M = 1, ..., 5$ and select the best model represents the training data.
Solution:

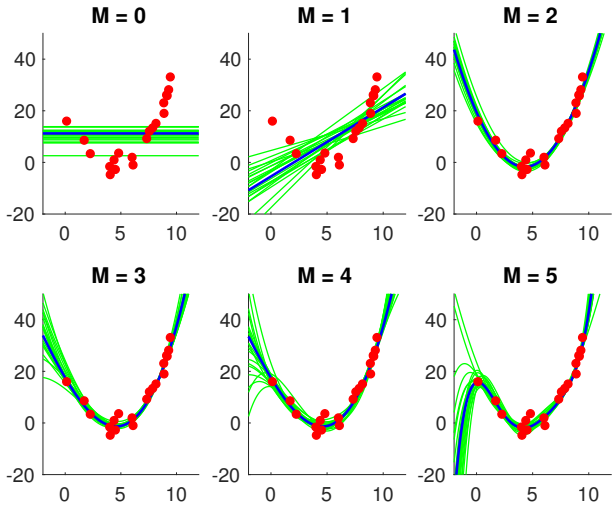

Figure 7: Bayesian model comparison for polynomials $M = 1, ..., 5$

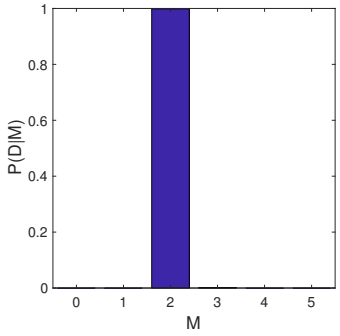

Figure 8: Model evidence for polynomials $M = 1, ..., 5$

We are looking for the order of the polynomial that maximizes the model evidence. Therefore, we select $M = 2$ that best represents the training data.

E. In Bayesian regression we should not regularize the bias term. This can be easily accomplished by using centered input ($\mathbf{\Phi}$) and output data ($\mathbf{y}$). The bias term can be computed and added to your solution through a post-processing operation. Using such a procedure, repeat the plot in B above for the optional model selected in B.

Solution:

From the notes, first, we use centered input ($\mathbf{y}$)

$y \to y - \bar{y}$

The X don't have the constant term. Then we do the same procedure as B. Then we can w. In order to get w_0,

$$w_0 = \bar{y} - \bar{X}^T w \tag{62}$$

From B, we find that $M = 4$ is the optimal model. Therefore, we repeat $M = 4$

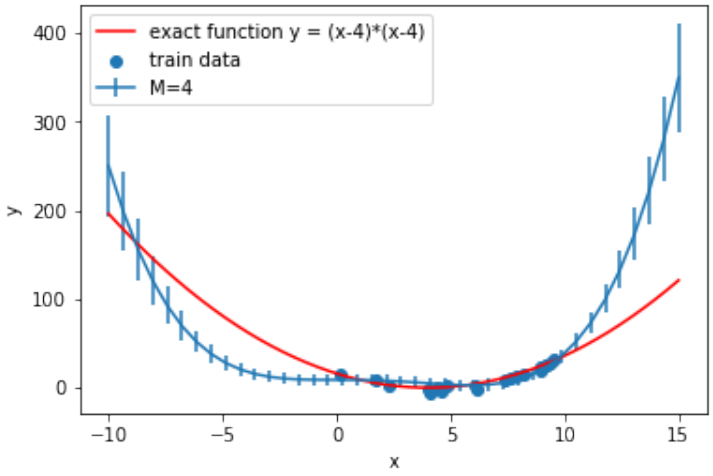

Figure 9: Predictive mean and the predictive error bars for polynomials $M = 4$